

2.9 Control Survey of the Terrestrial Inventory

Edgar Kaufmann, Andreas Schwyzer

To make mistakes is not desirable. To believe that no mistakes are made is naive. To repress mistakes is not credible. To not know the mistakes takes away the opportunity to improve upon them. To loose track of the mistakes can lead to wrong estimates and conclusions.

The control survey, which was conducted in the framework of the terrestrial survey for the second National Forest Inventory (NFI), served to detect errors and shortcomings as well as to quantify, to evaluate, and if possible, to eliminate them (see Chapter 2.8).

2.9.1 Purpose of the Control Survey

The amount of the deviation (random or systematic) of different assessments on the same objects, which were independently conducted by two survey teams, allowed conclusions to be drawn about the reproducibility of the inventory results. The quality of both assessments was regarded to be equal.

Systematic differences that were brought about by the results from the first and second survey teams indicated differences in the survey procedure or in the survey conditions. Causes of such differences could have been faulty instruments, incorrect handling of instruments, wrong assessments or different survey dates.

The control surveys (or the second surveys) allowed for periodic checks of the measurements and assessments of the survey teams during the field survey and to improve them through training.

The second surveys revealed unclear definitions of the attributes, allowed for an evaluation of the quality for the measured attributes and indicated the potential improvement with respect to future surveys.

This chapter introduces the analysis methods employed and illustrates them with the help of a few examples.

2.9.2 Methods

Out of the 6,400 terrestrial sample plots of the second NFI, approximately 10% were randomly selected for a checkup by a second survey team. The chosen sample plots were not known to the first survey team. Both surveys were conducted independently from each other. The second survey also encompassed the full catalog of the terrestrially measured attributes.

During the field survey, training courses were organized for the field teams in six week intervals. During these courses the results of the comparison between the first and second survey of the previous six week period were presented and discussed together with the survey teams. The main focus of the training days was directed towards the shortcomings, which were discovered by comparing the first and the second surveys.

The analysis methods for the continuous and categorical data were different. The applied statistical methods and evaluation criteria for the data quality based on selected attributes are discussed in the following chapters.

2.9.2.1 Analysis of Continuous Attributes

Measured and counted values from the first and second survey teams were compared graphically with each other. With this, the precision of the survey was visually illustrated, particularly during the training of the survey teams (Chapter 2.8). The displayed graphs were scatter diagrams that showed differences of individual measurements (Fig. 1), bar diagrams, which

presented mean deviations and dispersions of deviations (Figure 3), and frequency distributions of deviations between the assessments of the survey and the control teams (Fig. 2).

For the continuous variables, the average difference between the measured values of the first and second survey teams is a measure for systematic differences of the measurements. The standard deviation of the differences (s_d) is a measure for the random measurement error.

Assuming that the first and second survey team are measuring with the same random error,

$\frac{s_d}{\sqrt{2}}$ is an estimate of the random measurement error.

The hypothesis to distinguish whether or not differences were systematic or random was tested with the help of the t-statistic. Since outliers strongly influence or distort the parametric statistical measure, all analyses were done in two ways: 1) with all of the data included and 2) with only a 99% quantile of the data included (Table 1), i.e. the percentage of data that had the largest deviation between the measurements of the survey team and the ones from the control team was not included in the reduced data set.

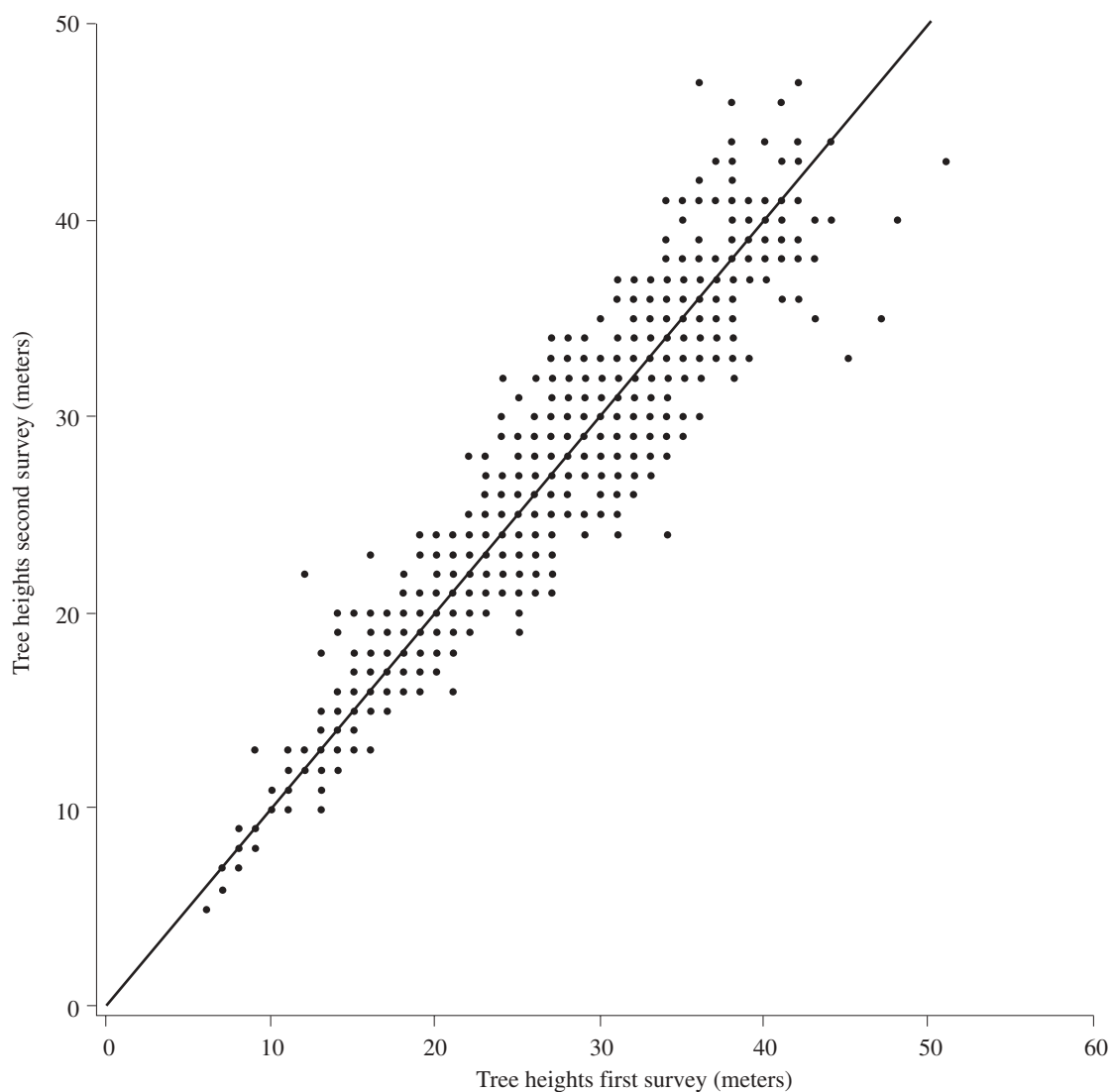


Figure 1. Tree height measurements of the first survey team and the second survey team.

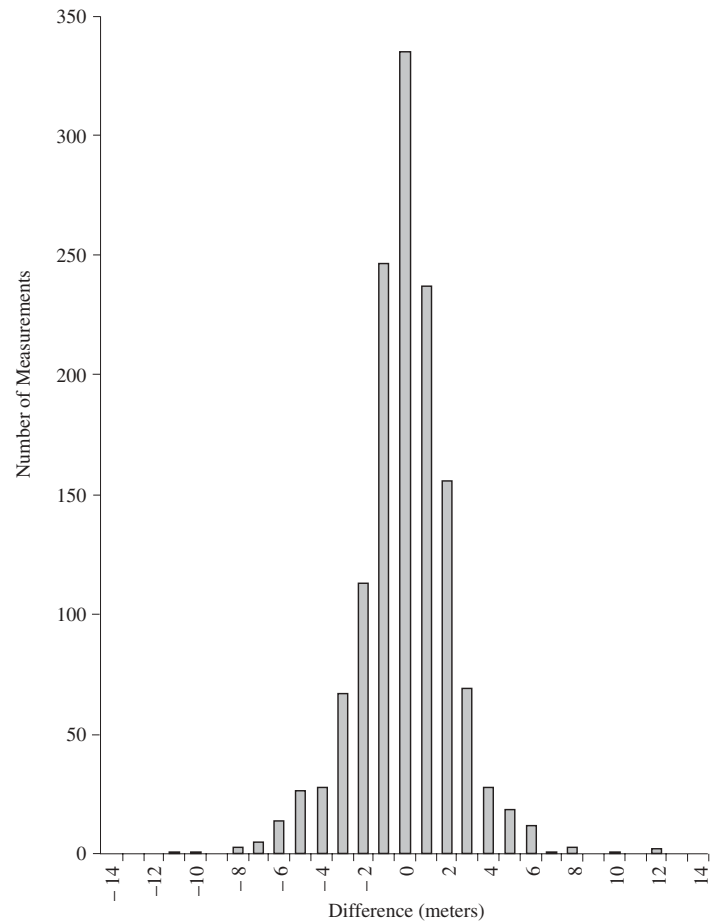


Figure 2. Tree height measurement differences between the first and second survey.

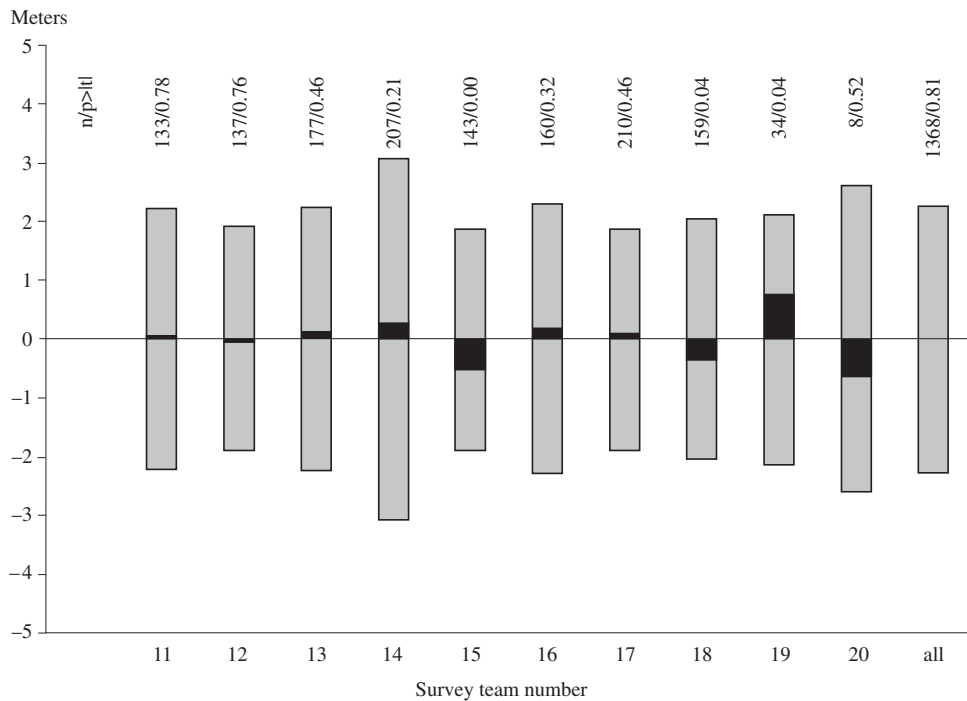


Figure 3. Measurement of tree height: Result of the check assessment in the second NFI by survey teams (group numbers 11–20). Average deviation ($1/n \sum(x_i - y_i)$, black) between the measurements of the first survey team (x_i) and those of the second survey team (y_i) at the same tree. Standard deviation of the measurement differences (hatch, difference $D_i = x_i - y_i$), N: Number of measurements, $p > |t|$: observed level of significance of the t-distribution.

Table 1. Statistical parameters of the **measurement data** in the second NFI, and the difference between first and second survey team.

All Data					
Attribute	Number	Systematic error	Random error	$P_{(T)}$	Result
DBH	8360	-0.8 mm	5.7 mm	0.0001	Significant difference (reason: growth)
D7	1236	-1.8 mm	13.1 mm	0.0007	Significant difference (reason: growth)
Tree height	1368	0.01 m	1.6 m	0.8115	Very good agreement
99%-Percentile					
Attribute	Number	Systematic error	Random error	$P_{(T)}$	Result
DBH	8294	-0.8	3.7 mm	0.0001	Even without outliers significant difference
D7	1224	-2.1 mm	11.5 mm	0.0001	Even without outliers significant difference
Tree height	1357	0.01 m	1.5 m	0.9282	Very good agreement
Only data from the dormant season					
Attribute	Number	Systematic error	Random error	$P_{(T)}$	Result
DBH	1319	-0.2 mm	3.9 mm	0.1189	Very good agreement
D7	225	-0.8 mm	11.3 mm	0.4302	Very good agreement
Tree height	239	0.26 m	1.3 m	0.0389	Moderate agreement

2.9.2.2 Analysis of Categorical Attributes

Most of the attributes assessed in the NFI were categorical. It was not enough to compare the evaluation of the first and second survey teams with each other in contingency tables, and to figure out the proportion of corresponding estimates. The problem is that the fewer classes a categorical attribute has, the larger the proportion of agreeing observations is if they are randomly distributed over all classes. An attribute with two classes is not considered as precise as an attribute with five classes if the proportion of corresponding estimates for both attributes is the same. Thus, suitable test statistics were chosen. These statistics allowed on one hand to compare the assessment qualities of the different attributes with each other. On the other hand, they are robust, i.e. the measures are valid even if numbers of cell frequencies are small and distributions are skewed. These kinds of statistics produce measures of association, i.e. they measure the tightness of the relationship between the assessments of the first and second survey teams. These measures helped to detect whether or not there was any asymmetry around the main diagonal of contingency tables. Also, the marginal distributions of the contingency tables were checked. It was examined if the cell frequencies of an attribute as a result of the assessments of the first and second survey teams were different, with no respect to concordances or discordances at the same objects. The test statistics used for nominal attributes were not the same as those used for ordinal attributes. For the test statistics the notation was as follows:

- x: Code for a categorical attribute that was determined by the first survey team
- y: Code for a categorical attribute that was determined by the second survey team
- l: Index for an observation

- i: Row index in a contingency table
- j: Column index in a contingency table
- k: Number of categories of an attribute

- n: Number of observations
 n_{ij} : Number of observations in cell i, j
 H_0 : Null hypotheses: Different assessments by first and second survey teams are random.
 H_1 : Alternative hypotheses: Assessments by the first and second survey teams are systematically different.
 α : Error probability for accepting the alternative hypotheses (accepting the alternative hypotheses for $\alpha < 0.05$)

Testing the Assessment of Ordinal Attributes

The **Sign Test** measures the direction of deviations between two assessments in contingency tables (SACHS 1974; SAS 1990a; SIEGEL and CASTELLAN 1988).

Test statistic: $S = p - n/2$

where p : Number of pairs with $x_l - y_l > 0$
 n : Number of pairs with $x_l - y_l \neq 0$

$H_1 (P_S < \alpha)$: Discordant assessments are not trend free. That is, a systematic increase of frequencies in certain directions exist.

By calculating ranks, the **Wilcoxon Rank Sum Test** measures, apart from the direction, the amount of the discordance in contingency tables (SACHS 1974; SAS 1990a; SIEGEL and CASTELLAN 1988). Large discordances are weighted higher than smaller ones.

Test statistic: $RS = \sum r_l^+ - \frac{n(n+1)}{4}$

where r_l^+ : rank of $|x_l - y_l|$ for $x_l - y_l \neq 0$

$H_1 (P_{RS} < \alpha)$: Direction and/or amount of discordant assessments are not random.

Gamma is a measure of association, which measures tightness of correlation between two ordinal scaled variables (GOODMAN and KRUSKAL 1979; SAS 1990a; SIEGEL and CASTELLAN 1988). Gamma approaches 0 for independence, 1 for complete dependence and -1 for complete negative dependence. It is possible that the number of concordant observations is small, even though the correlation is high. This is the case when the first survey team chose systematically higher or lower values in all categories than the second survey team. The test statistic is:

$$\text{Gamma} = \frac{(P - Q)}{(P + Q)}$$

where $P = \sum_i \sum_j n_{ij} A_{ij}$ and $Q = \sum_i \sum_j n_{ij} D_{ij}$

$$A_{ij} = \sum_{k>i} \sum_{l>j} n_{kl} + \sum_{k<i} \sum_{l<j} n_{kl} :$$

For each cell the number of observations n_{kl} for which the first and second survey teams either classified higher or lower than the code value of the cell considered.

$$D_{ij} = \sum_{k>i} \sum_{l<j} n_{kl} + \sum_{k<i} \sum_{l>j} n_{kl} :$$

For each cell the number of observations for which the second survey team classified higher and the first survey team classified lower (or vice versa) than the code value of the cell considered.

For ordinal attributes with at least five categories, the marginal distributions were tested with the **Kolmogorov-Smirnov Test** (SAS 1990b). This test is normally used to test continuous distributions. According to (SIEGEL and CASTELLAN 1988), this test can also be used for ordinal data. A significant test statistic means that the frequency distribution of attribute values, as they were measured by the first and second survey teams on one attribute, must be regarded as different. It is possible, therefore, that both marginal distributions do not differ from each other, even for poor agreement of the assessments on the same object.

The test statistic D is the maximum difference between the relative cumulative distributions of the two independent frequency distributions, or the marginal distributions of the contingency tables respectively.

$$\text{Test statistic: } D = \max_{x=y} \left(\frac{F_1(x)}{n} - \frac{F_2(y)}{n} \right)$$

where $F_1(x) = \sum_{i|x_i \leq x}^n x_i$ and $F_2(y) = \sum_{j|y_j \leq y}^n y_j$: Cumulative frequencies of the marginal distributions

$H_1 (P_D < \alpha)$: The marginal distributions are different. That is, the first and second survey teams determined different frequencies of a certain attribute.

Nominal Attributes

For nominal data there is no rank order between the classes. It is meaningless in which order the categories are listed in a contingency table.

The **McNemar Test** is a special case of the Cochran-Mantel-Haenszel Statistics (AGRESTI 1990; 1990a; SAS 1990b; SIEGEL and CASTELLAN 1988) and a special case of the sign test. The measure indicates whether discordant classifications are randomly distributed within a table, or whether they are more frequent in certain cells. The test for a $k \times k$ contingency table developed by BOWKER (1948, cited in LIENERT 1962) is analogous to the McNemar test for a 2×2 Table.

$$\text{Test statistic: } CMH = \sum_{i>j} \frac{(n_{ij} - n_{ji})^2}{(n_{ij} + n_{ji})} \text{ with } \frac{k(k-1)}{2} \text{ degrees of freedom}$$

$H_1 (P_{CMH} < \alpha)$: There exists an asymmetry with respect to the main diagonal, i.e. not all frequencies in corresponding cells, which are in a symmetric position to the main diagonal, are the same. This means that the first and second survey teams did not describe the attribute the same way.

The association measure **Kappa** ($-1 \leq \text{Kappa} \leq 1$) measures the tightness of the relationship for nominal data, while considering the expected random agreement (AGRESTI 1996; SIEGEL and CASTELLAN 1988). Especially for very skewed distributions (most of the observations fall into one category), or for attributes with few categories the probability is high that two assessments match at random. Kappa is calculated in the following way:

$$\text{Test statistic } K = \frac{[P(A) - P(E)]}{[1 - P(E)]}$$

where $P(A)$: Proportion of agreeing observations

$P(E)$: Proportion of agreeing observations when no connection exists between two

ratings of the same object: $P(E) = \sum_{i=j=1}^k \pi_{i+} \pi_{+j}$

where $\pi_{i+} = \frac{n_{i+}}{n}$ and $\pi_{+j} = \frac{n_{+j}}{n}$ (+: all categories of a row or a column)

2.9.3 Evaluation of the Measurement and Assessment Accuracy

The application of the tests mentioned above are illustrated in the following by means of selected examples.

2.9.3.1 Continuous Data

Continuous data are usually more precisely recorded than categorical data which are based on judgments. In the NFI the following measuring quantities were recorded:

- Diameter at breast height ($d_{1.3}$) on trees with $12 \text{ cm} \leq d_{1.3} \leq 60 \text{ cm}$
- Circumference at breast height of trees with $d_{1.3} > 60 \text{ cm}$
- Diameter at 7 m height (d_7) of the tariff sample trees
- Tree height (H) of the tariff sample trees
- Number of trees per sample plot having a $d_{1.3} \geq 12 \text{ cm}$

The overall random measurement error for the $d_{1.3}$ was estimated at 5.7 mm (Table 1). The mean systematic difference between the measurements of the first and the second survey team was 0.8 mm. This small difference was statistically significant at the 95% level (t-test). The systematic difference can be explained for measurements during the growing season by the time gap between the first and the second measurements. Figure 4 shows how the time gap between the two measurements effected the systematic differences of the measurements that were taken during the growing season. The further the two measurements were apart, the larger the average measurement difference was. The cause was attributed to the diameter growth between the measurements. The average difference between the first and second survey was random only if those measurements were compared with each other that were taken after the annual diameter growth was finished (Table 1).

The estimated random measurement error of 13.1 mm for the d_7 was larger than the one for the $d_{1.3}$ (Table 1). Nevertheless, it was small considering the difficulties connected with the measurements. With respect to the significance of the systematic differences, the same was true as for the $d_{1.3}$.

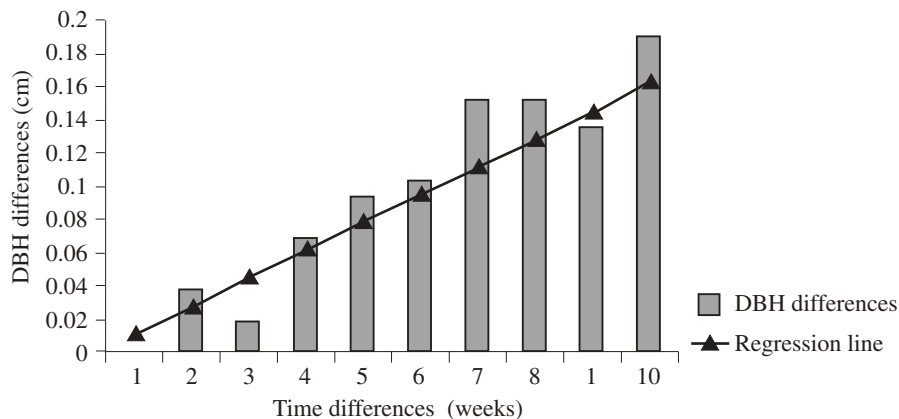


Figure 4. DBH differences between the first and second survey by time differences between the recording dates of first and second survey team during the growing season (April 1 to August 31).

The random measurement error for the tree height amounted to 2.3 m (Table 1). Large differences ($>7 \text{ m}$) between two measurements on the same tree were rare (0.8% of all measurements). On average, the tree height measurements of the survey team were not significantly different from the measurements taken by the control team ($P_1 > 0,05$). Training effects were clearly visible (Figure 5), especially for measurements such as the tree height, which required some training. Both the maximum differences and the standard deviation of the differences decreased in the course of one year.

The quality of the individual tree measurements can be described overall as very satisfactory. The good quality of individual tree measuring quantities was fundamental for avoiding systematic biases for individual tree volume and, thus, for growing stock and increment estimates (Chapter 3).

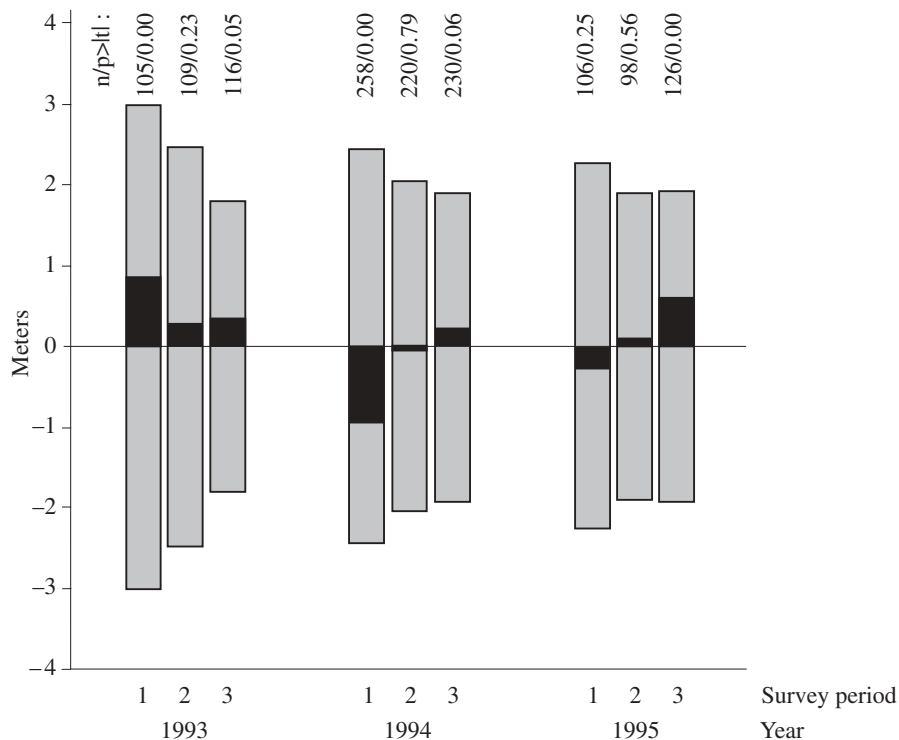


Figure 5. Measurement of tree heights: Results of the check assessment in the second NFI by survey period (1–3) within one year.

Average deviation ($\frac{1}{n}\sum(x_i - y_i)$, black) between the measurements of the first survey team (x_i) and those of the second survey team (y_i) at the same tree.

Standard deviation of the measurement differences (hatch, difference $D_i = x_i - y_i$),

N: Number of measurements, $p > |t|$: Observed level of significance of the t-distribution

2.9.3.2 Categorical Data

A large number of individual tree, stand, and site attributes in the NFI were assessed ocularly and were not based on measurements. Consistent training of the survey teams, clear assessment criteria that defined classification as precisely as possible, and good knowledge about forests by the survey teams are prerequisites for reliable and reproducible surveys. For the categorical data the following points must be kept in mind:

- Classification instructions are less precise than measurement instructions and always give the survey teams certain interpretation latitude.
- The interpretation latitude can lead to the preference of middle categories for ordinal variables. Good agreement between the ratings by the survey and the control teams can falsely indicate good reproducibility.
- Especially for binary variables (e.g., with the classes “present” and “not present”) with skewed distributions of the attribute (when most of the ratings fall into one class), a large proportion of matching ratings mean little for the rating accuracy. For ratings like this, the McNemar test, for example, is more suitable. This test measures the asymmetry of non-matching ratings independently of the number of matching ones.

Tables 2 and 3 show the number of observations (number of trees or sample plots) for the respective attributes that were rated by the first as well as the second survey teams.

Table 2. Contingency table and statistical parameter for **ordinal** attributes.

Social position		(Agreement: 76%, Gamma: 0.94, P _S : 0.00, P _{RS} : 0.00, P _D : 0.01)						
First survey	Code	Second survey						Total
		0	1	2	3	4	5	
Missing	0	1271	0	2	19	7	3	1302
Predominant	1	0	33	32	22	4	0	91
Dominant	2	10	45	655	388	7	0	1105
Co-dominant	3	25	18	603	4508	316	0	5470
Subdominant	4	17	1	4	345	1151	52	1570
Suppressed	5	5	0	2	4	223	72	306
Total		1328	97	1298	5286	1708	127	9844

Development stages		(Agreement: 64%, Gamma: 0.89, P _S : 0.04 P _{RS} : 0.03, P _D : 0.84)							
First survey	Code	Second survey							Total
		0	1	2	3	4	5	6	
Missing	0	9	2	0	0	0	0	0	11
Young growth / thicket	1	2	44	4	1	1	1	3	56
Pole wood	2	1	4	119	10	1	2	11	148
Young timber	3	1	2	22	70	19	2	9	125
Medium timber	4	1	0	1	24	101	24	21	172
Old timber	5	1	5	2	0	28	87	10	133
Mixed	6	1	5	13	10	20	12	59	120
Total		16	62	161	115	170	128	113	765

Mixture proportion		(Agreement: 82%, g: 0.94, P _S : 0.80, P _{RS} : 0.74, P _D : 1.00)					
First survey	Code	Second survey					Total
		0	1	2	3	4	
Missing	0	3	0	1	1	0	5
91–100 % Conifers	1	3	314	19	5	3	344
51–90% Conifers	2	0	31	95	14	2	142
11– 50% Conifers	3	0	5	15	54	22	96
0–10% Conifers	4	0	2	3	12	151	168
Total		6	352	133	86	178	755

Urgency of next operation		(Agreement: 36%, Gamma: 0.33, P _S : 0.00, P _{RS} : 0.00, P _D : 0.01)						
First survey	Code	Second survey						Total
		0	1	2	3	4	5	
Missing	0	63	6	20	8	16	5	118
Immediately	1	5	26	55	24	14	1	125
In 2 to 5 years	2	4	24	74	60	21	2	185
In 6 to 10 years	3	12	8	44	76	37	3	180
In 11 to 20 years	4	18	5	30	33	34	11	131
In >20 years	5	6	0	2	0	6	2	16
Total		108	69	225	201	128	24	755

Table 3. Contingency table and statistical parameter for **nominal** attributes.

Stand structure		Agreement: 65%, Kappa: 0.39, P _{CMH} : 0.76					
First survey	Code	Second survey					Total
		0	1	2	3	4	
Missing	0	3	2	0	0	0	5
Single layered	1	2	186	85	7	2	282
Multi-layered	2	1	91	278	42	3	415
Structured	3	0	3	12	21	2	38
Cluster structure	4	0	3	6	5	1	15
Total		6	285	381	75	8	755

Stand edge		Agreement: 77%, Kappa: 0.50, P _{CMH} : 0.28			
First survey	Code	Second survey			Total
		0	1	2	
Missing	0	3	0	2	5
Edge exists	1	3	177	74	254
No stand edge	2	0	95	401	496
Total		6	272	477	755

Traces of erosion		Agreement: 87%, Kappa: 0.36, P _{CMH} : 0.00				
First survey	Code	Second survey				Total
		1	2	3	4	
Channel	1	15	3	1	3	22
Surface	2	4	6	1	10	21
Slopes	3	4	2	3	6	15
None	4	29	18	14	636	697
Total		52	29	19	655	755

Geomorphological object		Agreement: 71%, Kappa: 0.57, P _{CMH} : 0.03									
First survey	Code	Second survey									Total
		1	2	3	4	5	6	7	8	9	
None	1	349	11	3	22	14	1	0	8	19	427
Scree	2	4	11	2	0	2	0	0	1	2	22
Loose rock	3	2	3	22	11	3	0	0	1	2	44
Boulder	4	5	0	14	44	13	0	0	1	0	77
Ledge of rock > 3m ²	5	8	1	4	10	78	0	0	7	2	110
Karst	6	0	0	1	0	0	1	0	0	0	2
Pit	7	1	0	0	0	0	0	2	0	0	3
Ravine	8	3	0	0	1	3	0	0	9	2	18
Trench over 80 cm	9	15	0	1	4	5	0	0	4	23	52
Total		387	26	47	92	118	2	2	31	50	755

The **social position** was rated differently by the two survey teams, as shown by the statistical measures in Table 2 ($P_S < 0,05$, $P_{RS} < 0,05$, $P_D < 0,05$). The group with the classes “predominant,” “dominant,” and “co-dominant” could be well separated from the group “subdominant” and “suppressed.” This fact was confirmed by the results of a correspondence analysis. Clear assignments within these two groups proved to be very difficult.

The **development stage** was sometimes not clearly determinable, especially when the stand boundary was close to a sample plot center. Despite this, the assessments on individual sample plots conducted by both teams turned out to be not clearly different. The value of the sign test ($P_S = 0,04$) and the rank sum test ($P_{RS} = 0,03$) were right on the borderline. The two marginal distributions were not systematically different from each other ($P_D > 0,05$), meaning that the different stages of development were rated just as frequently by the first survey team as by the second team. The statistical measures here refer to the ordinal part of the table (code 1–5).

The stand **mixture proportions** were well assessed. The correlation between the ratings of the first and second survey teams was very large ($\text{Gamma} = 0.94$). Furthermore, both the related assessments of individual objects, as well as the marginal distributions, were not significantly different from each other ($P_S > 0,05$; $P_{RS} > 0,05$; $P_D > 0,05$).

The **urgency of next silvicultural treatment**, however, could not be objectively assessed. The assessment of this attribute reflected the subjective opinion of the experts as indicated by weak correlations with a $\text{gamma} = 0.33$, systematically different classifications by first and second survey teams ($P_S < 0,05$, $P_{RS} < 0,05$), and different marginal distributions ($P_D < 0,05$).

The measures for the assessment of the **stand structure** and **stand boundary** in Table 3 show a low correlation between the first and second surveys ($\text{Kappa} = 0.39$ and $\text{Kappa} = 0.5$). However, there is no significant asymmetry with respect to the main diagonal ($P_{CMH} > 0,05$) in the contingency tables.

The low correlation for the attributes “**traces of erosion**” ($\text{Kappa} = 0.36$) and “**geomorphological objects**” ($\text{Kappa} = 0.57$) and, at the same time, the large proportion of matching observations (87% and 71%) was mainly due to the fact that such traces and objects were not found on most of the sample plots. These attributes were systematically evaluated differently by the first and second survey teams ($P_{CMH} < 0.05$).

If no asymmetry was found in the contingency table, and the marginal distribution of the first survey was not different from that of the second one, it was reasonable to assume that the frequency distribution of an attribute was assessed correctly. Forest areas identified with certain attribute values were in these cases assumed to be reliable, even if the assessments of the individual object had poor agreement.

Systematic error, however, can arise if poorly reproducible attributes are combined with other attributes, either for stratification (e.g. growing stock stratified by stand structure) or for attribute derivations (see Chapter 4.4.). Large random differences between the assessment of an attribute by the first and second survey teams result in ineffective stratification by this attribute. The use of poorly reproducible attributes for the derivation of other attributes is dubious. The plotwise or treewise combination of such an attribute with another attribute is also questionable.

The quality of assessments should not be judged based only on the test statistics, but always in connection with the contingency tables, especially with respect to frequencies of individual attribute values.

2.9.4 Outlook

The methods presented here were used to periodically analyze all variables during the terrestrial survey (KAUFMANN 1995). Additional studies are necessary in order to uncover the cause of misjudgments and to improve the survey quality for future inventories.

2.9.5 Literature

- AGRESTI, A., 1990: *Categorical Data Analysis*. New York: Wiley. 558 p.
- AGRESTI, A., 1996: *An Introduction to Categorical Data Analysis*. New York, Wiley. 290 pp.
- GOODMAN, L.A.; KRUSKAL, W.H., 1979. *Measures of Association for Cross Classifications*. New York, Berlin, Springer. 146 pp.
- KAUFMANN, E., 1995: *Kontrollaufnahmen LFI2, Übersicht 1993, 1994, 1995*. Birmensdorf, Eidgenössische Forschungsanstalt WSL, LFI, internal reports.
- LIENERT, G.A., 1962: *Verteilungsfreie Methoden in der Biostatistik*. Meisenheisen am Glan, Anton Hain KG. 360 pp.
- SACHS, L., 1974: *Angewandte Statistik*. 4 ed. Berlin, Springer. 545 pp.
- SAS, 1990a: *Procedures Guide*. Cary, NC, USA, SAS Institute Inc.
- SAS, 1990b: *SAS/STAT User's Guide*. Cary, NC, USA, SAS Institute Inc.
- SIEGEL, S.; CASTELLAN, N.J., 1988: *Nonparametric Statistics for behavioral sciences*. New York, McGraw-Hill. 399 pp.